A Simple Homotopy Algorithm for Compressive Sensing

Lijun Zhang[*] **Tianbao Yang**[†] **Rong Jin**[‡] **Zhi-Hua Zhou**[*]
[*]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[†]Department of Computer Science, the University of Iowa, Iowa City, USA
[‡]Department of Computer Science and Engineering, Michigan State University, East Lansing, USA
[‡]Institute of Data Science and Technologies at Alibaba Group, Seattle, USA
{zhanglj, zhouzh}@lamda.nju.edu.cn tianbao-yang@uiowa.edu rongjin@cse.msu.edu

Abstract

In this paper, we consider the problem of recovering the s largest elements of an arbitrary vector from noisy measurements. Inspired by previous work, we develop an homotopy algorithm which solves the ℓ_1-regularized least square problem for a sequence of decreasing values of the regularization parameter. Compared to the previous method, our algorithm is more *efficient* in the sense it only updates the solution once for each intermediate problem, and more *practical* in the sense it has a simple stopping criterion by checking the sparsity of the intermediate solution. Theoretical analysis reveals that our method enjoys a linear convergence rate in reducing the recovery error. Furthermore, our guarantee for recovering the top s elements of the target vector is tighter than previous results, and that for recovering the target vector itself matches the state of the art in compressive sensing.

1 Introduction

Compressive Sensing (CS) is a new paradigm of data acquisition that enables reconstruction of sparse or compressible signals from a relatively small number of linear measurements (Candès and Tao, 2006; Donoho, 2006). The standard assumption is that one has access to linear measurements of the form

$$\mathbf{y} = U^\top \mathbf{x}_* + \mathbf{e}$$

where $\mathbf{x}_* \in \mathbb{R}^d$ is the *unknown* target vector, $U \in \mathbb{R}^{d\times m}$ is the sensing matrix and $\mathbf{e} \in \mathbb{R}^m$ is a vector of noise. Recently, substantial process has been made in designing the encoder U and the associated decoder Δ which recovers $\mathbf{x}_*$ from U and $\mathbf{y}$ (Davenport et al., 2012).

One of the most famous decoders for CS is the ℓ_1-regularized least squares (ℓ_1-LS) formulation, known as Lasso in statistics (Tibshirani, 1996), given by

$$\min \frac{1}{2}\|\mathbf{y} - U^\top \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1.$$

The recovery performance of Lasso has been extensively studied and generally speaking, it can recover $\mathbf{x}_*$ up to the noise level under appropriate assumptions (Daubechies et al., 2004; Tropp, 2006; Zhang, 2009).

Under the assumption that λ is given beforehand, Xiao and Zhang (2012) propose a proximal-gradient homotopy method to improve the efficiency of ℓ_1-LS. The key idea is to solve the ℓ_1-LS problem for a sequence of decreasing values of the regularization parameter, and use an approximate solution at the end of each stage to warm start the next stage. In this study, we make three steps further. We show that (i) it is sufficient to run composite gradient mapping only *once* for each regularization parameter, (ii) the target regularization parameter can be detected *adaptively* based on the sparsity of the current solution, and (iii) this simple algorithm can deliver a tighter bound for recovering the s largest elements of $\mathbf{x}_*$.

For a vector $\mathbf{x}$, we denote by $\mathbf{x}^s$ the vector that contains the s largest elements of $\mathbf{x}$. Let $\hat{\mathbf{x}}$ be the solution returned by our algorithm. Under the assumption that U is a sub-Gaussian random matrix, our algorithm is able to reduce the recovery error *exponentially* over iterations, and the final recover error $\|\hat{\mathbf{x}}-\mathbf{x}_*^s\|_2$ is smaller than

$$O\left(\sqrt{\frac{s\log d}{m}}\left(\|\mathbf{e}\|_2 + \|\mathbf{x}_* - \mathbf{x}_*^s\|_2\right) + \|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2\right)$$

where $\mathbf{e}$ is the vector of noise, and $\|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2$ is the

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

ℓ_2-norm of the s largest elements of $\mathbf{x}_* - \mathbf{x}_*^s$. In contrast, previous analysis in CS (Davenport et al., 2012) can only upper bound the recovery error for $\mathbf{x}_*^s$ by

$$O\left(\|\mathbf{e}\|_2 + \|\mathbf{x}_* - \mathbf{x}_*^s\|_2\right).$$

Thus, our recovery guarantee could be significantly better than previous results if the ℓ_2-norm of $\mathbf{x}_* - \mathbf{x}_*^s$ is not concentrated on its s largest elements (i.e., $\|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2 \ll \|\mathbf{x}_* - \mathbf{x}_*^s\|_2$). Following the triangle inequality, we obtain the following bound for recovering $\mathbf{x}_*$

$$\|\hat{\mathbf{x}} - \mathbf{x}_*\|_2 \leq O\left(\|\mathbf{x}_* - \mathbf{x}_*^s\|_2 + \sqrt{\frac{s\log d}{m}}\|\mathbf{e}\|_2\right)$$

which matches state of the art (DeVore et al., 2009).

2 Related Work

Existing algorithms in CS could be roughly categorized into convex optimization based approaches and greedy approaches (Davenport et al., 2012; Blumensath et al., 2012). Roughly speaking, convex approaches have better theoretical guarantee, while greedy approaches are more efficient.

In the noise-free setting, Candès and Tao (2005) pose the following ℓ_1-minimization problem, denoted by Δ_1, for decoding

$$\min \|\mathbf{x}\|_1 \quad \text{s.t. } U^\top \mathbf{x} = \mathbf{y}.$$

To analyze the recovery performance, they introduce the Restricted Isometry Property (RIP) for matrices. Define the isometry constant of U as the smallest number δ_s such that the following holds for all s-sparse vectors $\mathbf{x} \in \mathbb{R}^d$

$$(1-\delta_s)\|\mathbf{x}\|_2^2 \leq \|U^\top \mathbf{x}\|_2^2 \leq (1+\delta_s)\|\mathbf{x}\|_2^2.$$

It has been shown that if $\delta_s + \delta_{2s} + \delta_{3s} < 1$ or $\delta_{2s} < \sqrt{2}-1$, the decoder Δ_1 yields perfect recovery for *all* s-sparse vectors $\mathbf{x}_* \in \mathbb{R}^d$ (Candès and Tao, 2005; Candès, 2008). If U is constructed as the random matrix with independent sub-Gaussian columns, a sufficient condition is to take $m = \Omega(s\log d)$ measurements (Mendelson et al., 2008).

In the general case, Candès (2008) propose the following convex formulation, denoted by Δ_1^ϵ, for decoding

$$\min \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{y} - U^\top \mathbf{x}\|_2 \leq \epsilon$$

where ϵ is an upper bound of $\|\mathbf{e}\|_2$. Let $\Delta_1^\epsilon(U^\top \mathbf{x}_* + \mathbf{e})$ be the solution returned by the above decoder. Suppose U satisfies the RIP of order $2s$ with $\delta_{2s} < \sqrt{2}-1$, we have

$$\|\Delta_1^\epsilon(U^\top \mathbf{x}_* + \mathbf{e}) - \mathbf{x}_*\|_2 \leq O\left(\frac{\|\mathbf{x}_* - \mathbf{x}_*^s\|_1}{\sqrt{s}} + \epsilon\right)$$

for *all* $\mathbf{x}_* \in \mathbb{R}^d$. An obvious drawback of this approach is that we must have a good a priori estimate of $\|\mathbf{e}\|_2$. This limitation is soon addressed by Wojtaszczyk (2010), who shows that the decoder Δ_1 performs very well even in the noise setting. In particular, if U is a Gaussian random matrix and $m = \Omega(s\log d)$, with an overwhelming probability, we have

$$\|\Delta_1(U^\top \mathbf{x}_* + \mathbf{e}) - \mathbf{x}_*\|_2 \leq O\left(\frac{\|\mathbf{x}_* - \mathbf{x}_*^s\|_1}{\sqrt{s}} + \|\mathbf{e}\|_2\right)$$

for *all* $\mathbf{x}_* \in \mathbb{R}^d$. Another possible way is to estimate the noise level under the Bayesian framework (Ji et al., 2008).

Notice that in the above inequalities, the ℓ_2-norm of the recovery error is upper bounded by the ℓ_1-norm of the corresponding error of the best s-term approximation. To make the result more consistent, it is natural to ask whether we could upper bound $\|\Delta(U^\top \mathbf{x}_* + \mathbf{e}) - \mathbf{x}_*\|_2$ by $\|\mathbf{x}_* - \mathbf{x}_*^s\|_2$ for some decoder Δ. Unfortunately, even in the noise-free setting, if we want the following inequality

$$\|\Delta(U^\top \mathbf{x}_*) - \mathbf{x}_*\|_2 \leq O(\|\mathbf{x}_* - \mathbf{x}_*^s\|_2)$$

to hold for all $\mathbf{x}_* \in \mathbb{R}^d$, the number of measurements m needs to be on the order of d (Cohen et al., 2009). This difficulty motivates the study of *instance optimality in probability*, which means we are looking for some performance guarantee that holds with a high probability for an arbitrary but *fixed* vector $\mathbf{x}_*$. When U is a Gaussian random matrix and $m = \Omega(s\log d)$, Wojtaszczyk (2010) shows that with an overwhelming probability

$$\|\Delta_1(U^\top \mathbf{x}_* + \mathbf{e}) - \mathbf{x}_*\|_2 \leq O\left(\|\mathbf{x}_* - \mathbf{x}_*^s\|_2 + \|\mathbf{e}\|_2\right)$$

for any *fixed* $\mathbf{x}_* \in \mathbb{R}^d$. This result is extended to more general families of matrices by DeVore et al. (2009).

The typical greedy approaches include Matching Pursuit (MP) (Mallat and Zhang, 1993), Iterative Hard Thresholding (IHT) (Blumensath and Davies, 2008; Garg and Khandekar, 2009), and Compressive Sampling Matching Pursuit (CoSaMP) (Needell and Tropp, 2009). Due to space limitation, we only give a brief description of IHT. In each iteration, IHT first performs gradient descent with respect to $\|U^\top \mathbf{x} - \mathbf{y}\|_2^2$ and then applies hard-thresholding to keep the s largest elements, in contrast to the soft-thresholding in our method. Let $\mathbf{x}_t$ denote the solution of IHT in the t-th iteration. The updating rule is given by

$$\mathbf{x}_{t+1} = \left[\mathbf{x}_t - \eta U(U^\top \mathbf{x}_t - \mathbf{y})\right]^s$$

where η is the step size. Under certain RIP condition, it has been shown that the recovery error of IHT can

Algorithm 1 A Simple Homotopy Algorithm

Input: Sensing Matrix $U \in \mathbb{R}^{d\times n}$, Measurements $\mathbf{y} \in \mathbb{R}^m$, Shrinking Parameter γ, Sparsity s, Maximum Number of Iterations T,

Initialize $\mathbf{x}_1 = 0$, $\lambda_1 = \|U\mathbf{y}\|_\infty$
for $t = 1, 2, \ldots, T$ **do**
  $\mathbf{x}_{t+1} = P_{\lambda_t}\left(\mathbf{x}_t - U(U^\top \mathbf{x}_t - \mathbf{y})\right)$
  if $\|\mathbf{x}_{t+1}\|_0 > 2s$ **then**
    Return $\mathbf{x}_t$
  end if
  $\lambda_{t+1} = \gamma\lambda_t$
end for
Return $\mathbf{x}_{T+1}$

be upper bounded by

$$O\left(\|\mathbf{x}_* - \mathbf{x}_*^s\|_2 + \frac{\|\mathbf{x}_* - \mathbf{x}_*^s\|_1}{\sqrt{s}} + \|\mathbf{e}\|_2\right)$$

for *all* $\mathbf{x}_* \in \mathbb{R}^d$ (Blumensath et al., 2012).

3 A Simple Homotopy Algorithm for Compressive Sensing

We first introduce the proposed homotopy algorithm, and then present its theoretical guarantee.

3.1 The Algorithm

In our algorithm, we solve a sequence of ℓ_1-regularized least squares (ℓ_1-LS) with decreasing regularization parameters. In particular, we set

$$\lambda_1 = \|U\mathbf{y}\|_\infty, \quad \lambda_{t+1} = \gamma\lambda_t$$

for some $\gamma < 1$. For each intermediate ℓ_1-LS problem, we use the solution from the previous iteration as the initial point, and perform composite gradient mapping (Nesterov, 2013) to update it *once*, i.e.,

$$\begin{aligned}\mathbf{x}_{t+1} =& \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} \langle \mathbf{x}, U(U^\top \mathbf{x}_t - \mathbf{y})\rangle + \frac{\|\mathbf{x} - \mathbf{x}_t\|_2^2}{2} + \lambda_t\|\mathbf{x}\|_1 \\ =& P_{\lambda_t}\left(\mathbf{x}_t - U(U^\top \mathbf{x}_t - \mathbf{y})\right)\end{aligned}$$

where $P_{\lambda_t}()$ is the soft-thresholding operator (Donoho, 1995) defined as

$$P_{\lambda_t}(\alpha) = \begin{cases} 0, & \text{if } |\alpha| \le \lambda_t; \\ \text{sign}(\alpha)(|\alpha| - \lambda_t), & \text{otherwise.}\end{cases}$$

The algorithm will stop when the sparsity of intermediate solution exceeds the budget $2s$. The above procedure is summarized in Algorithm 1.

Although the basic idea of our algorithm is motivated from the proximal-gradient homotopy (PGH) method (Xiao and Zhang, 2012), we make the following substantial extensions.

- Multiple iterations are needed by PGH to solve the intermediate optimization problem to achieve some predefined precision, in contrast, our algorithm only updates *once* for each problem.
- While PGH assumes a target regularization parameter is given beforehand, our algorithm is able to detect it *adaptively* based on the sparsity of the intermediate solution.
- Xiao and Zhang (2012) only analyze the recovery error of PGH for exactly sparse vectors, in comparison, we provide recovery guarantee for an *arbitrary* vector. This difference is more fundamental in the context of compressive sensing.

3.2 Main Results

We first describe the assumptions about the sensing matrix

$$U = \frac{1}{\sqrt{m}}[\mathbf{u}_1, \ldots, \mathbf{u}_m] = \frac{1}{\sqrt{m}}[\mathbf{v}_1, \ldots, \mathbf{v}_d]^\top \in \mathbb{R}^{d\times m}$$

where $\mathbf{u}_i \in \mathbb{R}^d, i = 1, \ldots, m$, and $\mathbf{v}_j \in \mathbb{R}^m, j = 1, \ldots, d$. We assume

- Both the columns and rows of U are sub-Gaussian vectors, and for the sake of simplicity, we assume the sub-Gaussian norm is smaller than 1. That is, for any vector $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{z} \in \mathbb{R}^m$, we have $\|\mathbf{x}^\top \mathbf{u}_i\|_{\psi_2} \le \|\mathbf{x}\|_2, i = 1, \ldots, m$ and $\|\mathbf{z}^\top \mathbf{v}_j\|_{\psi_2} \le \|\mathbf{z}\|_2, j = 1, \ldots, d$, where $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm of random variables (Vershynin, 2012).
- $\mathbf{u}_i$'s are sampled independently from an isotropic distribution, that is, $\mathrm{E}[\mathbf{u}_i\mathbf{u}_i^\top] = I$.

Examples of such random vectors are Gaussian vectors and Rademacher vectors.

Given a fixed vector $\mathbf{x}_*$, we receive measurements $\mathbf{y} = U^\top \mathbf{x}_* + \mathbf{e}$, where $\mathbf{e}$ encodes potential noise. In this paper, $\mathbf{e}$ is *unknown* and we make no assumption about the noise model. Our goal is to recover the s largest elements of $\mathbf{x}_*$ from U and $\mathbf{y}$.

We have the following theorem to bound the recovery error.[1]

Theorem 1. *Let $\hat{\mathbf{x}}$ be the solution output from our algorithm and T is the maximum number of iterations allowed. Choosing $\gamma = \frac{1+\sqrt{2}}{3}$, then, with a probability at least $1 - 6Te^{-\tau}$, we have*

$$\|\hat{\mathbf{x}} - \mathbf{x}_*^s\|_2 \le \max\left(6\Lambda, 3\lambda_1\sqrt{s}\gamma^T\right)$$

[1] As a reminder, $\mathbf{x}^s$ denotes the vector that contains the s largest elements of $\mathbf{x}$.

where

$$\Lambda = \sqrt{\frac{s(\tau+\log d)}{m}}\|\mathbf{e}\|_2 + C\left(\|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2 + \sqrt{\frac{\tau + s\log(d/s)}{m}}\|\mathbf{x}_* - \mathbf{x}_*^s\|_2\right) \tag{1}$$

for some universal constant C, provided

$$C\sqrt{\frac{\tau + s\log(d/s)}{m}} \leq \frac{1}{6}. \tag{2}$$

Remark The constants in the above theorem should not be treated literally, because we have made no effort to optimize them. Generally speaking, a smaller γ will lead to a fast convergence rate, but a larger constant, which is 6 now, in the recovery guarantee.

The above theorem implies that the recovery error reduces *exponentially* until it reaches $O(\Lambda)$. Thus, with a sufficiently large T, the recovery error of $\mathbf{x}_*^s$ can be upper bounded by

$$O\left(\sqrt{\frac{s\log d}{m}}\left(\|\mathbf{e}\|_2 + \|\mathbf{x}_* - \mathbf{x}_*^s\|_2\right) + \|(\mathbf{x}_* - \mathbf{x}_*^s)^s\|_2\right).$$

In the noise-free setting, our analysis also implies exact recovery of s-sparse vectors, since $\Lambda = 0$ if $\|\mathbf{e}\|_2 = 0$ and $\|\mathbf{x}_*\|_0 \leq s$.

From the literature of CS (Davenport et al., 2012, Theorem 1.14), we find that previous analysis is able to upper bound the recovery error of $\mathbf{x}_*^s$ by $c\left(\|\mathbf{x}_* - \mathbf{x}_*^s\|_2 + \|\mathbf{e}\|_2\right)$ for some constant $c > 1$. Thus, when m is large enough, our upper bound could be significantly smaller than the existing results. Finally, it is worth to mention that if our goal is to recover $\mathbf{x}_*$, following the triangle inequality, our analysis yields the same upper bound as previous studies (DeVore et al., 2009).

3.3 A Post-processing Step

Since $\hat{\mathbf{x}}$ is $2s$-sparse and $\mathbf{x}_*^s$ is s-sparse, one may ask whether it is possible to find a good s-sparse vector to approximate $\mathbf{x}_*^s$. The following theorem shows that we can simply select the s largest elements of $\hat{\mathbf{x}}$ to approximate $\mathbf{x}_*^s$ and the recovery error is on the same order.

Theorem 2. *Let $\mathbf{y} \in \mathbb{R}^d$ be a s-sparse vector. Then, we have*

$$\|\mathbf{x}^s - \mathbf{y}\|_2 \leq \sqrt{3}\|\mathbf{x} - \mathbf{y}\|_2, \ \forall \mathbf{x} \in \mathbb{R}^d.$$

4 Analysis

We here present the proofs of main theorems. The omitted proofs are provided in the supplementary material.

4.1 Proof of Theorem 1

Notice that starting our algorithm with $\lambda_1 = \|U\mathbf{y}\|_\infty$ has the same effect as staring with $\|U\mathbf{y}\|_\infty\gamma^{-k}$, $k \in \mathbb{Z}$, which means we can set λ_1 as large as we need. Thus, without loss of generality, we can assume

$$\lambda_1 \geq \frac{1}{3\sqrt{s}}\max\left(\|\mathbf{x}_*^s\|_2, 6\Lambda\right). \tag{3}$$

We first state two theorems that are central to our analysis. Theorem 3 reveals that the recovery error of our algorithm will reduce by a constant factor until it reaches the optimal level. Then, Theorem 4 shows that the recovery error will remain small, as long as the sparsity of the solution does not exceed $2s$.

We denote by $\mathcal{S}_*$ and $\mathcal{S}_t$ the support set of $\mathbf{x}_*^s$ and $\mathbf{x}_t$, respectively.

Theorem 3. *Assume $|\mathcal{S}_t \setminus \mathcal{S}_*| \leq s$, $\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \leq 3\lambda_t\sqrt{s}$, and $\Lambda \leq \frac{1}{2}\lambda_t\sqrt{s}$, where Λ is given in (1). Then, with a probability at least $1 - 6e^{-\tau}$, we have*

$$|\mathcal{S}_{t+1} \setminus \mathcal{S}_*| \leq s \text{ and } \|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \leq 3\lambda_{t+1}\sqrt{s}$$

provided the condition in (2) is true.

Theorem 4. *Assume $|\mathcal{S}_t| \leq 2s$, $\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \leq 6\Lambda$, and $\Lambda > \frac{1}{2}\lambda_t\sqrt{s}$, where Λ is given in (1). If $|\mathcal{S}_{t+1}| \leq 2s$, then with a probability at least $1 - 6e^{-\tau}$, we have*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \leq 2(1+\sqrt{3})\Lambda$$

provided the condition in (2) is true.

We continue the proof of Theorem 1 in the following. Let

$$k = \min\left\{t : \Lambda > \frac{1}{2}\lambda_t\sqrt{s}\right\} \overset{(3)}{>} 1.$$

In the following, we will show that the recovery error $\|\mathbf{x}_t - \mathbf{x}_*^s\|_2$ will first decrease exponentially as t approaches k, and then keep below 6Λ.

$T < k$ From (3), we have $\|\mathbf{x}_1 - \mathbf{x}_*^s\|_2 = \|\mathbf{x}_*^s\|_2 \leq 3\lambda_1\sqrt{s}$. Since the condition $\Lambda \leq \frac{\lambda_t\sqrt{s}}{2}$ holds for $t = 1, \ldots, T$, we can apply Theorem 3 to bound the recovery error in each iteration. Thus, with a probability at least $1 - 6Te^{-\tau}$, we have

$$\|\hat{\mathbf{x}} - \mathbf{x}_*^s\|_2 = \|\mathbf{x}_{T+1} - \mathbf{x}_*^s\|_2 \leq 3\lambda_{T+1}\sqrt{s} = 3\lambda_1\sqrt{s}\gamma^T.$$

$T \geq k$ From the above analysis, with a probability at least $1-6(k-1)e^{-\tau}$, we have $\|\mathbf{x}_k - \mathbf{x}_*^s\|_2 \leq 3\sqrt{s}\lambda_k$ and $|\mathcal{S}_k \setminus \mathcal{S}_*| \leq s$, which also means our algorithm arrives the k-th iteration. In the k-th iteration, there will be two cases: $|\mathcal{S}_{k+1}| > 2s$ and $|\mathcal{S}_{k+1}| \leq 2s$. For the first case, our algorithm terminates, and return $\mathbf{x}_k$ as the final solution, implying

$$\|\hat{\mathbf{x}} - \mathbf{x}_*^s\|_2 = \|\mathbf{x}_k - \mathbf{x}_*^s\|_2 \leq 3\lambda_k\sqrt{s} \leq 6\Lambda.$$

For the second case, our algorithm keeps running, and we can bound the recovery error of $\mathbf{x}_{k+1}$ by Theorem 4. If $T = k$ or $|\mathcal{S}_{k+2}| > 2s$, our algorithm terminates and return $\mathbf{x}_{k+1}$ as the final solution, which implies

$$\|\hat{\mathbf{x}} - \mathbf{x}_*^s\|_2 = \|\mathbf{x}_{k+1} - \mathbf{x}_*^s\|_2 \leq 2(1+\sqrt{3})\Lambda.$$

Otherwise, our algorithm keeps running. Since $2(1+\sqrt{3}) \leq 6$, the condition in Theorem 4 are satisfied, and thus can be applied repeatedly to bound the recovery error for all the rest iterations.

4.2 Proof of Theorem 3

We need the following theorem to analyze the behavior of the composite gradient descent.

Theorem 5. *Suppose $\mathbf{x}_t - \mathbf{x}_*^s$ is a $3s$-sparse vector. With a probability at least $1-6e^{-\tau}$, we have*

$$\begin{aligned}&\left\|\left[U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right]^s\right\|_2 \\ \leq &\Lambda + C\sqrt{\frac{\tau + s\log(d/s)}{m}}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2\end{aligned}$$

where Λ is given in (1).

Given a set $\mathcal{S} \subseteq [d]$, $\mathbf{x}_\mathcal{S}$ denotes the vector which coincides with $\mathbf{x}$ on $\mathcal{S}$ and has zero coordinates outside $\mathcal{S}$. We denote the sub-gradient of $\|\cdot\|_1$ by $\partial\|\cdot\|_1$.

Using the fact that

$$\begin{aligned}&\mathbf{x}_{t+1} \\ =&\operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\left\|\mathbf{x} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right)\right\|_2^2 + \lambda_t\|\mathbf{x}\|_1,\end{aligned}$$

we have

$$\begin{aligned}&0 \\ \leq &\left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) + \lambda_t\partial\|\mathbf{x}_{t+1}\|_1,\right. \\ &\left.\mathbf{x}_*^s - \mathbf{x}_{t+1}\right\rangle \\ \leq &\left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right), \mathbf{x}_*^s - \mathbf{x}_{t+1}\right\rangle \\ &+ \lambda_t\|\mathbf{x}_*^s\|_1 - \lambda_t\|\mathbf{x}_{t+1}\|_1 \\ \leq &\left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right), \mathbf{x}_*^s - \mathbf{x}_{t+1}\right\rangle \\ &+ \lambda_t\|\mathbf{x}_*^s\|_1 - \lambda_t\|[\mathbf{x}_{t+1}]_{\mathcal{S}_*}\|_1 \\ \leq &\left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right), \mathbf{x}_*^s - \mathbf{x}_{t+1}\right\rangle \\ &+ \lambda_t\|[\mathbf{x}_{t+1} - \mathbf{x}_*^s]_{\mathcal{S}_*}\|_1 \\ \leq &\left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right), \mathbf{x}_*^s - \mathbf{x}_{t+1}\right\rangle \\ &+ \lambda_t\sqrt{s}\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2\end{aligned}$$

and thus

$$\begin{aligned}&\lambda_t\sqrt{s}\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \\ \geq &\left\langle \mathbf{x}_{t+1} - \mathbf{x}_t + U\left(U^\top\mathbf{x}_t - \mathbf{y}\right), \mathbf{x}_{t+1} - \mathbf{x}_*^s\right\rangle \\ = &\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2^2 \\ &+ (\mathbf{x}_{t+1} - \mathbf{x}_*^s)^\top\left(U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right).\end{aligned} \tag{4}$$

According to Theorem 5, with a probability at least $1-6e^{-\tau}$, we have

$$\begin{aligned}&\left\|\left[U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right]^s\right\|_2 \\ \leq &\Lambda + C\sqrt{\frac{\tau + s\log(d/s)}{m}}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \\ \overset{(2)}{\leq} &\Lambda + \frac{1}{6}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \leq \lambda_t\sqrt{s}.\end{aligned}$$

The above inequality implies the magnitude of the smallest $d-s$ elements of $\mathbf{x}_t - U(U^\top\mathbf{x}_t - \mathbf{y}) - \mathbf{x}_*^s$ is smaller than λ_t. Combining with the fact that

$$\mathbf{x}_{t+1} = P_{\lambda_t}\left(\mathbf{x}_t - U(U^\top\mathbf{x}_t - \mathbf{y})\right),$$

it is easy to verify that $|\mathcal{S}_{t+1} \setminus \mathcal{S}_*| \leq s$. Furthermore,

$$\begin{aligned}&\left|(\mathbf{x}_{t+1} - \mathbf{x}_*^s)^\top\left(U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right)\right| \\ \leq &\left(\|[\mathbf{x}_{t+1} - \mathbf{x}_*^s]_{\mathcal{S}_*}\|_2 + \|[\mathbf{x}_{t+1} - \mathbf{x}_*^s]_{\mathcal{S}_{t+1}\setminus\mathcal{S}_*}\|_2\right)\lambda_t\sqrt{s} \\ \leq &\lambda_t\sqrt{2s}\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2.\end{aligned}$$

Substituting the above inequality into (4), with a probability at least $1-6e^{-\tau}$, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2^2 \leq \left(\lambda_t\sqrt{s} + \lambda_t\sqrt{2s}\right)\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2$$

and thus

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2 \leq (1+\sqrt{2})\lambda_t\sqrt{s} \leq 3\lambda_{t+1}\sqrt{s}.$$

4.3 Proof of Theorem 4

We need to reuse (4) that appears in the analysis of Theorem 3. According to Theorem 5, with a probability at least $1-6e^{-\tau}$, we have

$$\begin{aligned}&\left\|\left[U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right]^s\right\|_2 \\ \overset{(2)}{\leq} &\Lambda + \frac{1}{6}\|\mathbf{x}_t - \mathbf{x}_*^s\|_2 \leq 2\Lambda.\end{aligned}$$

Notice that $\mathbf{x}_{t+1} - \mathbf{x}_*^s$ is $3s$-sparse in this case, and it is easy to verify that

$$\begin{aligned}&\left|(\mathbf{x}_{t+1} - \mathbf{x}_*^s)^\top\left(U\left(U^\top\mathbf{x}_t - \mathbf{y}\right) - (\mathbf{x}_t - \mathbf{x}_*^s)\right)\right| \\ \leq &2\sqrt{3}\Lambda\|\mathbf{x}_{t+1} - \mathbf{x}_*^s\|_2.\end{aligned}$$

Substituting the above inequality into (4), we have

$$\begin{aligned}\|\mathbf{x}_{t+1}-\mathbf{x}_*^s\|_2^2 \leq& \left(\lambda_t\sqrt{s}+2\sqrt{3}\Lambda\right)\|\mathbf{x}_{t+1}-\mathbf{x}_*^s\|_2\\ \leq& 2(1+\sqrt{3})\Lambda\|\mathbf{x}_{t+1}-\mathbf{x}_*^s\|_2,\end{aligned}$$

and thus

$$\|\mathbf{x}_{t+1}-\mathbf{x}_*^s\|_2 \leq 2(1+\sqrt{3})\Lambda.$$

4.4 Proof of Theorem 5

In the analysis, we need to bound $\|(UU^\top\mathbf{z})^s\|_2$ for a *fixed* vector $\mathbf{z}\in\mathbb{R}^d$, and $\|[(UU^\top-I)\mathbf{z}]^s\|_2$, for *all* $3s$-sparse vectors $\mathbf{z}\in\mathbb{R}^d$. Thus, we build the following two theorems.

Theorem 6. *For a fixed $\mathbf{z}\in\mathbb{R}^d$, with a probability at least $1-2e^{-\tau}$, we have*

$$\left\|\left(UU^\top\mathbf{z}\right)^s\right\|_2 \leq C\left(\sqrt{\frac{\tau+s\log(d/s)}{m}}\|\mathbf{z}\|_2+\|\mathbf{z}^s\|_2\right)$$

for some constant $C>0$.

Theorem 7. *With a probability at least $1-2e^{-\tau}$, for all $\mathbf{z}\in\mathbb{R}^d$ with $\|\mathbf{z}\|_0\leq 3s$, we have*

$$\left\|\left[(UU^\top-I)\mathbf{z}\right]^s\right\|_2 \leq C\sqrt{\frac{\tau+s\log(d/s)}{m}}\|\mathbf{z}\|_2$$

for some constant $C>0$.

We rewrite $U\left(U^\top\mathbf{x}_t-\mathbf{y}\right)-(\mathbf{x}_t-\mathbf{x}_*^s)$ as

$$\begin{aligned}&U\left(U^\top\mathbf{x}_t-\mathbf{y}\right)-(\mathbf{x}_t-\mathbf{x}_*^s)\\ =&U\left(U^\top\mathbf{x}_t-U^\top\mathbf{x}_*-\mathbf{e}\right)-(\mathbf{x}_t-\mathbf{x}_*^s)\\ =&\underbrace{UU^\top(\mathbf{x}_*^s-\mathbf{x}_*)}_{:=\mathbf{w}_a}+\underbrace{(UU^\top-I)(\mathbf{x}_t-\mathbf{x}_*^s)}_{:=\mathbf{w}_b}-\underbrace{U\mathbf{e}}_{:=\mathbf{w}_c}.\end{aligned}$$

Then, we have

$$\begin{aligned}&\left\|\left[U\left(U^\top\mathbf{x}_t-\mathbf{y}\right)-(\mathbf{x}_t-\mathbf{x}_*^s)\right]^s\right\|_2\\ \leq&\|\mathbf{w}_a^s\|_2+\|\mathbf{w}_b^s\|_2+\|\mathbf{w}_c^s\|_2.\end{aligned}$$

Bounding $\|\mathbf{w}_a^s\|_2$ According to Theorem 6, with a probability at least $1-2e^{-\tau}$, we have

$$\begin{aligned}&\|\mathbf{w}_a^s\|_2\\ =&\left\|\left[UU^\top(\mathbf{x}_*^s-\mathbf{x}_*)\right]^s\right\|_2\\ \leq&C\left(\sqrt{\frac{\tau+s\log(d/s)}{m}}\|\mathbf{x}_*-\mathbf{x}_*^s\|_2+\|(\mathbf{x}_*-\mathbf{x}_*^s)^s\|_2\right)\end{aligned}$$

for some constant $C>0$.

Bounding $\|\mathbf{w}_b^s\|_2$ Notice that $\mathbf{x}_t-\mathbf{x}_*^s$ is a $3s$-sparse vector. According to Theorem 7, with a probability at least $1-2e^{-\tau}$, we have

$$\begin{aligned}\|\mathbf{w}_b^s\|_2=&\left\|\left[(UU^\top-I)(\mathbf{x}_t-\mathbf{x}_*^s)\right]^s\right\|_2\\ \leq&C\sqrt{\frac{\tau+s\log(d/s)}{m}}\|\mathbf{x}_t-\mathbf{x}_*^s\|_2\end{aligned}$$

for some constant $C>0$.

Bounding $\|\mathbf{w}_c^s\|_2$ Since $U=\frac{1}{\sqrt{m}}[\mathbf{v}_1,\ldots,\mathbf{v}_d]^\top$, and we assume $\mathbf{v}_i$ is a sub-Gaussian vector. We have

$$\|\mathbf{v}_j^\top\mathbf{e}\|_{\psi_2}\leq\|\mathbf{e}\|_2,\ j=1,\ldots,d.$$

Using the property of Orliz norm (Koltchinskii, 2009, 2011), with a probability at least $1-2e^{-\tau}$, we have

$$|\mathbf{v}_j^\top\mathbf{e}|\leq\|\mathbf{v}_j^\top\mathbf{e}\|_{\psi_2}\sqrt{\tau}\leq\|\mathbf{e}\|_2\sqrt{\tau}.$$

By taking the union bound, we have, with a probability at least $1-2e^{-\tau}$,

$$\|U\mathbf{e}\|_\infty=\frac{1}{\sqrt{m}}\max_j|\mathbf{v}_j^\top\mathbf{e}|\leq\|\mathbf{e}\|_2\sqrt{\frac{\tau+\log d}{m}}$$

implying

$$\|\mathbf{w}_c^s\|_2=\|(U\mathbf{e})^s\|_2\leq\|\mathbf{e}\|_2\sqrt{\frac{s(\tau+\log d)}{m}}.$$

We complete the proof by combining the bounds for $\|\mathbf{w}_a^s\|_2$, $\|\mathbf{w}_b^s\|_2$, and $\|\mathbf{w}_c^s\|_2$.

4.5 Proof of Theorem 6

We define the set of s-sparse vectors with length smaller than 1 as

$$\mathcal{K}_{d,s}=\left\{\mathbf{x}\in\mathbb{R}^d:\|\mathbf{x}\|_2\leq 1,\|\mathbf{x}\|_0\leq s\right\}.$$

Then, it is easy to check that

$$\mathcal{E}_s(\mathbf{z}):=\max_{\mathbf{w}\in\mathcal{K}_{d,s}}\mathbf{w}^\top UU^\top\mathbf{z}=\|(UU^\top\mathbf{z})^s\|_2.$$

Let $\mathcal{K}_{d,s}(\epsilon)$ be a proper ϵ-net for $\mathcal{K}_{d,s}$ with the smallest cardinality, and $N(\mathcal{K}_{d,s},\epsilon)=|\mathcal{K}_{d,s}(\epsilon)|$ be the covering number for $\mathcal{K}_{d,s}$. We have the following lemma for bounding $N(\mathcal{K}_{d,s},\epsilon)$ (Plan and Vershynin, 2013, Lemma 3.3).

Lemma 1. *For $\epsilon\in(0,1)$ and $s\leq d$, we have*

$$\log N(\mathcal{K}_{d,s},\epsilon)\leq s\log\left(\frac{9d}{\epsilon s}\right).$$

Using the ϵ-net $\mathcal{K}_{d,s}(\epsilon)$, we define a discretized version of $\mathcal{E}_s(\mathbf{z})$ as

$$\mathcal{E}_s(\mathbf{z},\epsilon) = \max_{\mathbf{w}\in\mathcal{K}_{d,s}(\epsilon)} \mathbf{w}^\top UU^\top \mathbf{z}.$$

The following lemma relates $\mathcal{E}_s(\mathbf{z})$ with $\mathcal{E}_s(\mathbf{z},\epsilon)$.

Lemma 2. *For $\epsilon \in (0, 1/\sqrt{2})$, we have*

$$\mathcal{E}_s(\mathbf{z}) \leq \frac{\mathcal{E}_s(\mathbf{z},\epsilon)}{1-\sqrt{2}\epsilon}.$$

Based on the conclusion from Lemma 2, it is sufficient to bound $\mathcal{E}_s(\mathbf{z},\epsilon)$. To this end, we need the following lemma to bound the difference between $\mathbf{w}^\top UU^\top \mathbf{z} - \mathbf{w}^\top \mathbf{z}$ for a fixed $\mathbf{w}$.

Lemma 3. *For fixed $\mathbf{w}$ and $\mathbf{z}$ with $\|\mathbf{w}\|_2 \leq 1$, with a probability at least $1-2e^{-\tau}$, we have*

$$|\mathbf{w}^\top UU^\top \mathbf{z} - \mathbf{w}^\top \mathbf{z}| \leq C\sqrt{\frac{\tau}{m}}\|\mathbf{z}\|_2$$

for some constant $C > 0$.

By taking the union bound, with a probability at least $1-2e^{-\tau}$, we have

$$\begin{aligned}&\max_{\mathbf{w}\in\mathcal{K}_{d,s}(\epsilon)} \left|\mathbf{w}^\top UU^\top \mathbf{z} - \mathbf{w}^\top \mathbf{z}\right| \\ \leq & C\sqrt{\frac{\tau + s\log(9d/[\epsilon s])}{m}}\|\mathbf{z}\|_2.\end{aligned}$$

Since

$$\max_{\mathbf{w}\in\mathcal{K}_{d,s}(\epsilon)} \mathbf{w}^\top \mathbf{z} \leq \max_{\mathbf{w}\in\mathcal{K}_{d,s}} \mathbf{w}^\top \mathbf{z} = \|\mathbf{z}^s\|_2,$$

we have

$$\begin{aligned}\mathcal{E}_s(\mathbf{z},\epsilon) = & \max_{\mathbf{w}\in\mathcal{K}_{d,s}(\epsilon)} \mathbf{w}^\top UU^\top \mathbf{z} \\ \leq & \max_{\mathbf{w}\in\mathcal{K}_{d,s}(\epsilon)} \left|\mathbf{w}^\top UU^\top \mathbf{z} - \mathbf{w}^\top \mathbf{z}\right| + \max_{\mathbf{w}\in\mathcal{K}_{d,s}(\epsilon)} \mathbf{w}^\top \mathbf{z} \\ \leq & C\sqrt{\frac{\tau + s\log(9d/[\epsilon s])}{m}}\|\mathbf{z}\|_2 + \|\mathbf{z}^s\|_2.\end{aligned}$$

We complete the proof by using Lemma 2 with $\epsilon = 1/2$.

4.6 Proof of Theorem 7

Recall the definitions of $\mathcal{K}_{d,s}$ and $\mathcal{K}_{d,s}(\epsilon)$ in the proof of Theorem 6. Evidently, for any $\mathbf{z}$ with $\|\mathbf{z}\|_0 \leq 3s$, we have

$$\left\|\left[(UU^\top - I)\mathbf{z}\right]^s\right\|_2 \leq \mathcal{G}_{3s}\|\mathbf{z}\|_2$$

where

$$\mathcal{G}_{3s} = \max_{\mathbf{r}\in\mathcal{K}_{d,3s}} \left\|\left[(UU^\top - I)\mathbf{r}\right]^s\right\|_2.$$

Thus, what we need is to provide an upper bound for $\mathcal{G}_{3s}$.

Define

$$\mathcal{F}_s(\mathbf{r}) = \left\|\left[(UU^\top - I)\mathbf{r}\right]^s\right\|_2 = \max_{\mathbf{w}\in\mathcal{K}_{d,s}} \mathbf{w}^\top (UU^\top - I)\mathbf{r}.$$

We first provide an upper bound for $\mathcal{F}_s(\mathbf{r})$, and then $\mathcal{G}_{3s}$. From the analysis of Theorem 6, with a probability at least $1-2e^{-\tau}$, we have

$$\begin{aligned}\mathcal{F}_s(\mathbf{r},\epsilon) = & \max_{\mathbf{w}\in\mathcal{K}_{d,s}(\epsilon)} \mathbf{w}^\top (UU^\top - I)\mathbf{r} \\ \leq & C\sqrt{\frac{\tau + s\log(9d/[\epsilon s])}{m}}\|\mathbf{r}\|_2.\end{aligned}$$

for some constant $C > 0$. Following the proof of Lemma 2, it is straightforward to show that

$$\mathcal{F}_s(\mathbf{r}) \leq \frac{\mathcal{F}_s(\mathbf{r},\epsilon)}{1-\sqrt{2}\epsilon},\ \forall \epsilon \in (0, 1/\sqrt{2}).$$

Similar to the proof of Theorem 6, we can set $\epsilon = 1/2$. Thus, for a fixed $\mathbf{r}$, with a probability at least $1-2e^{-\tau}$, we have

$$\mathcal{F}_s(\mathbf{r}) \leq C\sqrt{\frac{\tau + s\log(d/s)}{m}}\|\mathbf{r}\|_2$$

for some constant $C > 0$.

Next, we repeat the above argument again to bound $\mathcal{G}_{3s}$. We define a discreted version of $\mathcal{G}_{3s}$ as

$$\mathcal{G}_{3s}(\epsilon) = \max_{\mathbf{r}\in\mathcal{K}_{d,3s}(\epsilon)} \mathcal{F}_s(\mathbf{r}).$$

By taking the union bound, with a probability at least $1-2e^{-\tau}$, we have

$$\mathcal{G}_{3s}(\epsilon) \leq C\sqrt{\frac{\tau + s\log(d/s) + 3s\log(9d/[3\epsilon s])}{m}}$$

for some constant $C > 0$. We complete the proof by using

$$\mathcal{G}_{3s} \leq \frac{\mathcal{G}_{3s}(\epsilon)}{1-\sqrt{2}\epsilon},\ \forall \epsilon \in (0, 1/\sqrt{2}).$$

5 Empirical Study

In this section, we perform several experiments to examine the recovery performance of our method.

We first examine the ability of exact recovery under noise-free setting. $\mathbf{x}_* \in \mathbb{R}^{10000}$ is a 20-sparse vector, that is generated randomly and normalized to unit length. We construct the sensing matrix $U \in \mathbb{R}^{10000\times 1000}$ as the Gaussian random matrix. Fig. 1 shows how the recovery error of our algorithm decreases with different choices of γ. It is clear that our algorithm achieves a linear convergence rate, and a smaller γ leads to a faster convergence rate.

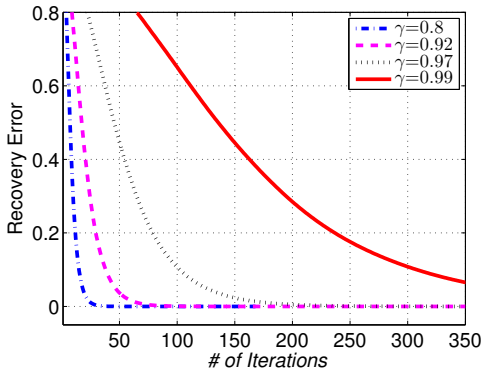

Figure 1: $\mathbf{x}_*$ is sparse.

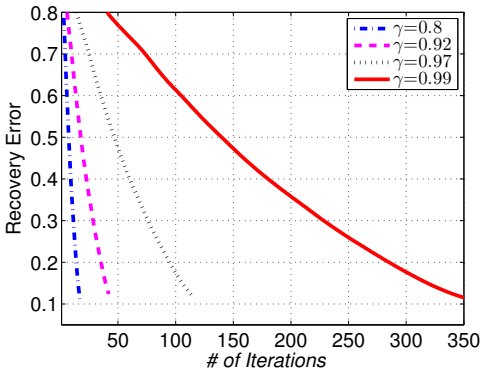

Figure 2: $\mathbf{x}_*$ is dense.

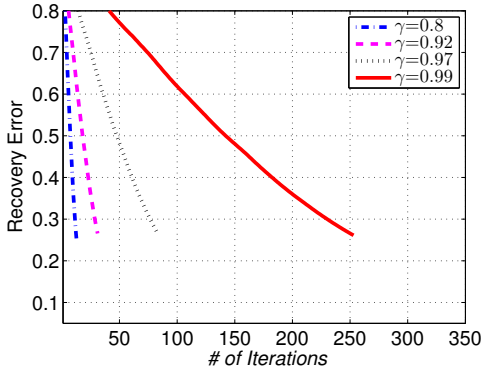

Figure 3: $\mathbf{x}_*$ is dense and $\mathbf{y}$ contains noise.

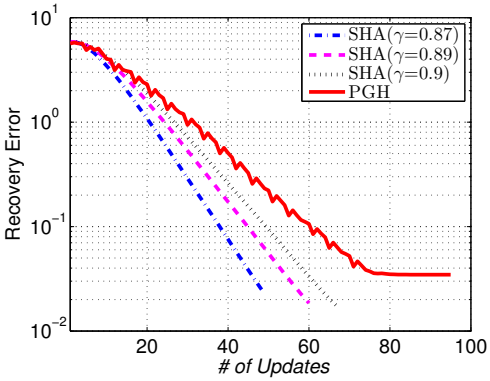

Figure 4: SHA versus PGH.

We then generate a unit length $\mathbf{x}_* \in \mathbb{R}^{10000}$ that has a power-law decay. Specifically, the i-th largest elements of $\mathbf{x}_*$ is proportional to i^{-1}. Fig. 2 shows the recovery error for $\mathbf{x}_*^{20}$ without noise and Fig. 3 shows the recovery error when the measurements are contaminated by Gaussian noise with $\|\mathbf{e}\|_2 = 0.5$. We observe that the recovery error still decreases rapidly until it reaches certain precision, at which the sparsity of the current solution exceeds 40 and our method terminates.

Finally, we compare our simple homotopy algorithm (SHA) with the proximal-gradient homotopy method (PGH). Following the setting in (Xiao and Zhang, 2012), we construct $U \in \mathbb{R}^{5000\times 1000}$ where entries are sampled independently from $\mathcal{U}(-\sqrt{3}, \sqrt{3})$, i.e., the uniform distribution over $[-\sqrt{3}, \sqrt{3}]$, $\mathbf{x}_* \in \mathbb{R}^{5000}$ where $\|\mathbf{x}_*\|_0 = 100$ and the non-zero entries are sampled from $\mathcal{U}(-1, 1)$, and $\mathbf{e} \in \mathbb{R}^{1000}$ where entries are sampled from $\mathcal{U}(-0.01, 0.01)$. Fig. 4 shows that our algorithm converges faster.

6 Conclusion and Future Work

In this paper, we provide a simple homotopy algorithm for CS that is more efficient and practical than its counterpart. Theoretical analysis shows that our algorithm has a linear convergence rate in reducing the recovery error, and the recovery guarantee for $\mathbf{x}_*^s$ could be much tighter than previous results under appropriate conditions.

Notice that existing studies in CS provide matching lower and upper bounds for recovering the whole vector $\mathbf{x}_*$ (Donoho, 2006; Cohen et al., 2009). It is unclear to us what would be the lower bound if our goal is to recover $\mathbf{x}_*^s$. We will investigate this issue in future. We will also study how to apply our algorithm to one-bit compressive sensing (Boufounos and Baraniuk, 2008; Plan and Vershynin, 2013; Zhang et al., 2014), which is a new setting of CS where the measurement is quantized to a single bit.

Acknowledgements

This research was supported by National Science Foundation of China (61333014), Collaborative Innovation Center of Novel Software Technology and Industrialization, NSF (IIS-1251031) and ONR Award N000141210431.

References

T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008.

T. Blumensath, M. E. Davies, and G. Rilling. Greedy algorithms for compressed sensing. In *Compressed Sensing, Theory and Applications*, chapter 8, pages 348–393. Cambridge University Press, 2012.

P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *Proceedings of the 42nd Annual Conference on Information Sciences and Systems*, pages 16–21, 2008.

E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9–10):589–592, 2008.

E. J. Candès and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22(1):211C–231, 2009.

I. Daubechies, M. Defrise, and C. D. Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok. Introduction to compressed sensing. In *Compressed Sensing, Theory and Applications*, chapter 1, pages 1–64. Cambridge University Press, 2012.

R. DeVore, G. Petrova, and P. Wojtaszczyk. Instance-optimality in probability with an ℓ_1-minimization decoder. *Applied and Computational Harmonic Analysis*, 27(3):275–288, 2009.

D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

R. Garg and R. Khandekar. Gradient descent with sparsification: An iterative algorithm for sparse recovery with restricted isometry property. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 337–344, 2009.

S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, 2008.

V. Koltchinskii. The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26 (3):301–321, 2009.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140 (1), 2013.

Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

J. A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing, Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

P. Wojtaszczyk. Stability and instance optimality for gaussian measurements in compressed sensing. *Foundations of Computational Mathematics*, 10(1), 2010. ISSN 1615-3375.

L. Xiao and T. Zhang. A proximal-gradient homotopy method for the ℓ_1-regularized least-squares problem. In *Proceedings of the 29th International Conference on Machine Learning*, pages 839–846, 2012.

L. Zhang, J. Yi, and R. Jin. Efficient algorithms for robust one-bit compressive sensing. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, 2009.